The Low-Rank Decomposition of Correlation-Enhanced Superpixels for Video Segmentation

Haixia Xu[1,2], Edwin R. Hancock[2], Wei Zhou[1]

Abstract Low-rank decomposition (LRD) is an effective scheme to explore the affinity among superpixels in the image and video segmentation. However, the superpixel feature collected based on colour, shape, and texture may be rough, incompatible, and even conflicting if multiple features extracted in various manners are vectored and stacked straight together. It poses poor correlation, inconsistence on intra-category super-pixels, similarities on inter-category super-pixels. This paper proposes a correlation-enhanced superpixel for video segmentation in the framework of low-rank decomposition (LRD). Our algorithm mainly consists of two steps, feature analysis to establish the initial affinity among super-pixels, followed by construction of a correlation-enhanced superpixel. This work is very helpful to perform LRD effectively and find the affinity accurately and quickly. Experiments conducted on Datasets validate the proposed method. Comparisons with the state-of-the-art algorithms show higher speed and more precise in video segmentation.

Keywords Video segmentation · LRD · The enhanced super-pixel

0. Introduction

In recent years, there has been received increasing attraction on video segmentation in such vision applications as robot navigation, scene understanding, active recognition, etc. Various works on video segmentation have been introduced ranging from graph-based processing and spectral clustering in Felzenszwalb et al.(2004), Grundmann et al.(2010), Shi et al.(2000) Wang et al.(2011), Achanta et al.(2010), online streaming segmentation in Galasso et al(2012), Xu et al. (2012), to superpixel-based low rank optimization in Zhang et al. (2013), Cheng et al.(2011),Wang et al.(2012),Li et al. (2016), Li et al. (2016). And some benchmarks in Xu et al (2012), Galasso et al. (2013) have been developed to evaluate the progress of these methods.

Segmentation, this is, to assign consistent label to pixels with similar property. Video segmentation is to group pixels into several semantically consistent spatiotemporal parts over a video volume. Principally, the substantial work involves in mining the affinity among pixels (region) in the image or video. Superpixel-based image segmentation Zhang et al. (2013), Cheng et al.(2011), Li et al.(2016) and video segmentation Li et al.(2016) utilize low rank optimization to find the affinity among super-pixels. This is to say, the low rankness of the super-pixel feature matrix is the basis on which to find the affinity across super-pixels. To mine the affinity using LRD is to exploit the correlation among SPs. LRD is an optimization method, and the efficiency and speed of the optimization lie on correlation of columns (rows) of matrix. The lower the rank of a matrix is, the more efficient the LRD optimization is. In turn, the more accurate the segmentation is.

We extract multi-kinds of features, RGB color, HSV color, HOG, MOV. Ideally, superpixels in the identical semantic region have high correlation, even they have equal representation of vectors. In fact, when we measure whether the i^{th} SP is correlative with j^{th} SP, maybe the answer is that they are high correlation based on RGB color, and that they are low correlation based on

Communicated by Haixia Xu

Haixia Xu
xhxia2002@126.ocm

Edwin.R. Hancock
edwin.hancock@york.ac.uk

Wei Zhou
zhou_wei@xtu.edu.cn

1. Key Laboratory of Intelligent Computing and Information Processing, Ministry of Education, College of Information Engineering, Xiangtan University, Xiangtan, 411105, China

2. Department of Computer Science, University of York, York, YO10 5DD, UK

MOV, and so on, ie., the superpixel features collected based on colour, shape, and texture are incompatible, conflicting if multiple features extracted in various manners are vectored and stacked straight together. It poses poor correlation, inconsistence on intra-category super-pixels, similarities on inter-category super-pixels.

Inspired by linear and correlation representation, this paper proposes a correlation-enhanced super-pixel for video segmentation in the framework of low-rank decomposition (LRD). Based on the LRD framework, the video is clipped into a certain length clips. For each clip, superpixels are got via a traditional segmentation method. We analyse and enhance the correlation of super-pixels. Then perform LRD on the correlation-enhanced superpixels to find the affinity across the superpixels.

1. Related work

Segmentation is fundamental work in computer vision. Some of the state-of-the-art approaches on the unsupervised segmentation are reviewed in this section. Surveying the related literatures, we roughly classify them into three-folds according the generation of the affinity.

First, pixel-wise affinity segmentation, it involves in finding pixels with similar perceptual appearance. Graph-based segmentation, Normalized Cuts (Ncut) and spectral clustering Felzenszwalb et al.(2004), Grundmann et al.(2010), Shi et al.(2000) Wang et al.(2011), SILC Achanta et al.(2010) find the affinity by computing the difference from pixels in 2D image. These methods are extensive to video domain, and they find pixels with similarly perceptual appearance and spatiotemporal continuity and group them in a video volume. Video segmentation can be done frame by frame, and can also be performed via stacking pixels of all frames together to process a big size image segmentation. However, there are massive pixels to be processed as image resolution and video length increase. It is too time-consuming and needs much more memory to store, even exceeds memory of PC. The abovementioned methods therefore are often used to get the super-pixel (SP) or super-voxel (SV). VSS (Video segmentation with superpixels) Galasso et al. (2012) generates SPs via Ncut, then achieves image segmentation based on SPs.

Second, researchers propose streaming and online video segmentation to overcome the limitations of memory and space, which is one of the bottlenecks to process massive pixels in a video volume. Xu et al (2012) emploited steaming hierarchical video segmentation. The use of steaming framework reduces the consumption of memory and space.

Third, recently, low-rank optimization is widely used to mine data correlation in saliency extraction, image classification and segmentation. Here, superpixels or super-voxels from images or video are optimized via low-rank to find their affinity, for short, superpixel-based affinity for segmentation. Liu et al. (2013) and Yin et al. (2016) detailed the low-rank representation of subspace. Zhang et al. (2013) investigated the low-rank sparse coding and demonstrated the low-rankness of super-pixel. Cheng et al.(2011), Wang et al.(2012), Li et al. (2016) proposed low-rank affinity pursuit for image segmentation, they used multi feature observation to descript the super-pixel, and exploited the affinity via low-rank optimization. Li et al. (2016) presented Sub-Optimal Low-Rank Decomposition for efficient video segmentation (SOLD), and formulated the affinity with three cues to enhance the accuracy of segmentation. The utilization of low-rank optimization alleviates the influence of the data noise.

Besides, deep learning (DL) based approaches achieve much in the visual task of image and video segmentation Pixel or Image descriptors are fed into a DL segmentation network. Segmentation results are got via training and inference. Originally, DL is used in the task of image classification, then object detection & localization, tracking, segmentation and so on. As for the typical work of image segmentation, we classify it into several folders in the view of the DL network structure. First, plain network based segmentation structure. Long et al. (2017) proposed FCN network that was adapted from contemporary classification networks into fully convolutional networks and that their learned representations were transferred by fine-tuning to the segmentation task. DeepLab Chen et al. (2018), includes DeepLab v1 to DeepLab v3+. Ronneberger et al. (2015) designed U-Net, like a “U” shape, which has a contracting path to capture the context paired with an expanding path that enables precise localization. Second, Residual network based segmentation structure. It is more difficult to train as neural networks deepen. He et al. (2018) presented ResNet, a residual network learning framework, to ease the training of networks. Many works are followed by

Residual network based segmentation structure. For an instance, DeepLab v2 developed by Chen et al. (2016) replaces Vgg16 with Resnet-101. The performance of mIoU increases by 2%. Third, Generative adversarial network (GAN)-based structure. GAN is an adversarial training processing. Segmentation with GAN Luc et al. (2016) reinforces spatial contiguity in the output label map to yield accuracy improvement. The combination of Superpixel in Farnoush et al. (2018), Das et al. (2018), He et al. (2015), CRF or GAN with DL is used to reinforce spatial contiguity. An inherent challenge is the trade-off between accuracy and computational cost although DL contributes much improvement.

Whether DL-based approaches or conventional methods, have devoted great efforts to pixels-wise label. Superpixels are more semantic than pixels. Substantially, which region superpixels are grouped into lies on their feature description and their correlation. This paper proposes a correlation-enhanced superpixel for video segmentation in the framework of low-rank decomposition (LRD).

The remainder of this paper is organized as follows. Section 2 discusses the principle of LRD for video segmentation. A correlation-enhanced superpixel for video segmentation is proposed in Section 3. In Section 4 experiments and discussion are given. Final section presents concluding remarks as well as future work.

2. LRD for video segmentation

A. *Superpixel*

The superpixel is used for a variety of applications, e.g., human pose estimation, semantic pixel label, 3D reconstruction from the image and video segmentation, for it has more semantic information than the pixel. Our work of video segmentation begins with not pixels but superpixels. For the sake of the trade-off of time and precision, we divide each image frame into around 200 homogeneous and coherent superpixels by using SLIC segmentation algorithm proposed by Achanta et al.(2010)

B. *Multiple feature extraction*

Feature extraction plays a vital role in the visual analysis task. A number of efficient feature extraction methods have been developed. Owing to these works, we assume that a set of appearance and motion features are extracted from the i^{th} superpixel and combined into one single d-dimensional feature vector x_i for the superpixel representation. All of feature vectors of n superpixels form the data matrix observation $X = [x_1, x_2, \dots\dots, x_n] \in R^{d\times n}$.

C. *The Low-rank decomposition(LRD)*

Low-rank representation (LRR) in Liu et al. (2013), Liu et al. (2012) proposes a low-rank based criterion for subspace clustering. We assume that super-pixels belonging to the same semantic patch are derived from one identical low-rank subspace, and all superpixels in a certain spatiotemporal window lie on a union of multiple subspaces. According to the LRR Liu et al. (2013), the super-pixel feature observations have

$$X = XZ + E, \quad s.t. \quad rank(Z) \le r \tag{1}$$

where $Z \in R^{n\times n}$ is the desired low-rank affinity matrix among super-pixels, $E \in R^{d\times n}$ is the sparse corrupted noise. r denotes the low rank of the affinity matrix Z. Thus, the low-rank representation is modelled as a nuclear norm minimization problem

$$\min_{Z,E} \|Z\|_* + \lambda\|E\|_1, s.t.\ X = XZ + E \tag{2}$$

where $\|\cdot\|_*$, $\|\cdot\|_1$ denote the nuclear norm and L_1-norm of a matrix, respectively. λ is the balance factor of two terms.

Fixed rank-low representation (FRR) proposed in Liu et al. (2012) states that Z with fixed rank is more convenient to optimize and better to represent the structure of Z. As the fixed-rank constraint is imposed on Z, we explicitly express Z, non-uniquely, as a matrix product $Z=UV$, where $U \in R^{n\times r}$ and $V \in R^{r\times n}$, $r < \min(n, d)$. Meanwhile, to speed up optimization, FRR is to minimize the Frobenius norm instead of the nuclear norm of the representation Z as in the LRR.

$$\min_{Z,U,V,E} \|Z - UV\|_F^2 + \lambda\|E\|_1, s.t.\ X = XZ + E \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

C. *Video Segmentation with LRD*

Steaming framework in Xu et al (2012) is utilized to processing an arbitrary long video. The video is divided into overlapping clips in temporal window, and

segmented successively while enforcing consistency. The spectral segmentation in Shi et al. (2000) is used to obtain the final result.

3. The proposed method

A. Superpixel Feature analysis

Segmentation is to group pixels or superpixels with similar attributes into a region. The feature representation of the superpixel is the basis to group superpixels and to achieve the good performance. It is crucial to construct the discriminative and powerful feature for each superpixel. Feature selection and analysis are given below.

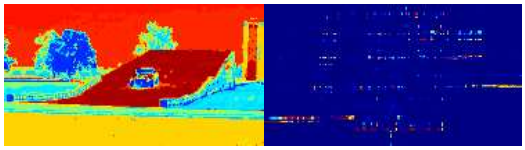
For each superpixel, feature descriptors are chosen based on colour, shape, texture, etc., so that they capture the discriminative characteristics while showing robustness to some noise, insensitive to illumination variation etc., as follows.

I. The colour histograms in the colour space of RGB, HSV acting as the appearance descriptor

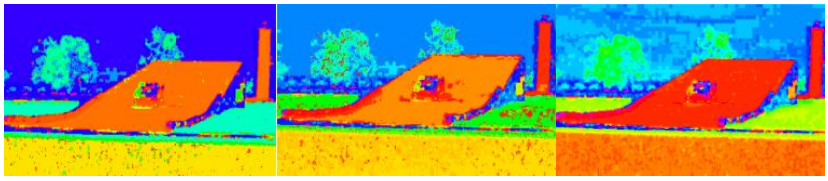
The colour is one of the most importance appearance attributes. We extract colour feature from the each superpixel and express it in the form of histogram vector. The uniform quantization is applied to each channel of RGB. Generally, the quantized bins n_{rgb} is set to 8, 12. We find that bins 12 is not much better than bins 8, and both bins 12 and bins 8 are supposed to illumination. To reduce the cost of computation, $n_{rgb} = 8$ is chosen to express the RGB colour feature of SP in Fig.1 (a). Note that, for the histogram matrix formed from all of super-pixels, the row of all 0's is deleted.

In order to add colour and reduce the impact of illumination change, HSV space is taken into account. The analysis is conducted in the manner of uniform and non-uniform quantization, respectively. Components H S V are uniformly quantized to bins 256, and Components H, S, V are non-uniformly quantized to bins 16, 8, 4, 3 in the different experiments. For instance, component H with bins 16 and component S with bins 4 form Qhs with bins n_{hsv} =64. H with bins 8, S with bins 3 and V with bins 3 form Qhsv with $n_{hsv} = 72$.

The combination of H, S, V is given in Fig.1 (b) by varying bins of H, S,V. It is seen from Fig.1(b) that uniform quantization in right side is not fit, and non-uniform quantization in the left side and middle are feasible. As a matter of fact, Qhs64 in the left is better than Qhsv72 in the middle. For Qhsv72, the presence of illumination component V poses bad impact on consistency region. The reduction of component V, even just the use of components H, S, will alleviate the impact of shadow and illumination.



(a) The uniform quantization in RGB space



(b) The quantization in HSV space

Fig.1 The quantization in color space. (a) the uniform quantization in RGB space. The left side is the quantized image, the right side is the histogram of the quantized image that is counted in each superpixel. (b) The quantization in HSV space. Non-uniform quantization Qhs64 in the left side, Qhsv72 in the middle, and uniform quantization Q256 in the right side.

II. The histogram of oriented gradients (HOG) indicating the shape, contour

HOG is relatively invariant to local geometric and photometric Transformations. We extract HOG to indicate the contour of each segment according to the method Dalal et al. (2005). Edge detection operator, for

the instance, Sobel filter, is used to obtain the gradient of image $I(u,v)$. The gradient amplitude $\nabla I(u,v)=\sqrt{(\frac{\partial I(u,v)}{\partial u})^2+(\frac{\partial I(u,v)}{\partial v})^2}$ and the orientation $\theta(u,v)=\arctan(\frac{\partial I(u,v)/\partial v}{\partial I(u,v)/\partial u})$ are calculated from each pixel. The orientation is divided into bins n_{hog} and each bin is counted by the weighted vote of gradient magnitude to form an orientation-based histogram. In this paper, in order to eliminate the impact of illumination change, the weight of each bin is improved by the relative gradient magnitude, i.e., the ratio of the gradient magnitude to the image intensity $I_r(u,v)=\frac{|\nabla I(u,v)|}{I(u,v)}$. Then, HOG with orientations bins n_{hog} are counted within each superpixel. Empirically, it is found that histogram channels with $n_{hog}=32$ perform best in our video segmentation experiments.

III. The histogram of optical flow (HOF) revealing the motion attribute between frames.

The HOF descriptor is a good fit to sequence frames. The motion is recorded between two concessive frames, which is useful to eliminate the clustering of moving object and background. Here, a simple filter [1,-1] is used to compute the temporal derivation for each two consecutive frames. Similar as for HOG, the magnitude and the orientation of spatial motion are obtained, in turn, the histogram of oriented motion gradient is formed in Duta et al. (2016). It is found empirically that histogram with $n_{hof}=32$ perform well in eliminating the clustering of moving object and background.

Each kind of descriptor, expressed in the form of histogram vector, is normalized by L_2-norm to avoid the impact of the size of superpixel, and then be concatenated into a long descriptor vector with dimension $d=n_{rgb}+n_{hsv}+n_{hog}+n_{hof}$.

B. The correlation-enhanced Super-pixel

Pixel-wise LRD optimization is robust to outers and noise. However, histogram-wise LRD needs to be tackled carefully. Multiple features of superpixels in the identical semantic area are not always coherent when they are used to discriminate patterns. The anisotropy kinds of features from the identical superpixel pose the unreliability, inaccuracy on the affinity *Z*. i.e., the multiple features may be compatible, or may be conflicting when they are inputted into LRD optimization.

Intra-category superpixels are drawn from the identical subspace, and all super-pixels in a certain spatiotemporal window lie on a union of subspaces. A linear combination of superpixels that are from one subspace still lies on this subspace, so the superpixel added linearly by other superpixels from the identical subspace is equivalent to the one increasing the correlative component. The superpixel enhanced by the correlative component will show high intra-category similarities whereas inter-category dissimilarities.

Inspired by the linear representation and correlation theory, we propose a correlation-enhanced super-pixel used for video segmentation. The details are below.

Firstly, we introduce a pre-defined affinity matrix W between superpixel *i* and superpixel *j*, which is used to analyse roughly the discrimination of the histogram descriptor. Since the feature is in the form of histogram vector, the affinity W is measured by Bhattacharrya coefficient Konstantinos (2008).

$$W_{ij}=\sum_{k=0}^{d}\sqrt{x_i(k)x_j(k)}\quad i=1,\dots,n,\quad j=1,\dots,n \tag{4}$$

where W_{ij} is the $(i,j)^{th}$ entry of the matrix W. Binarize W by threshold τ, $W_{ij}=1$ if $W_{ij}>\tau$. That is, vector x_i, x_j are correlative and belong to one semantic region with high probability. Otherwise $W_{ij}=0$.

Then, we enhance the correlation of intra-category superpixels by increasing the correlative component which is indicated by the above matrix W. The correlation-enhanced superpixel observations X_{new} are given by

$$X_{new}=X[(1-a)I_n+aW] \tag{5}$$

where a is set to 0~1, controlling the magnitude of the linear combination of superpixels. I_n is the n x n identity matrix. Specially,

$$\begin{cases}X_{new}=XW & if\ a=1\\ X_{new}=X & if\ a=0\end{cases}$$

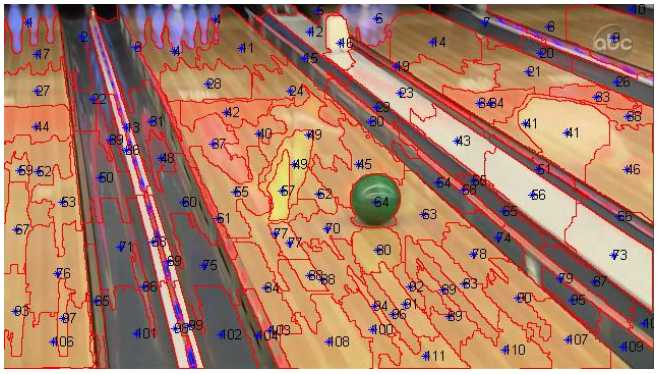
i.e., they are common superpixels when a=0.

The original superpixel and the correlation-enhanced superpixel are visualized in Fig.2, superpixels partitioned using SLIC in Fig.2 (a), and the original supepixel feature from the identical semantic region in

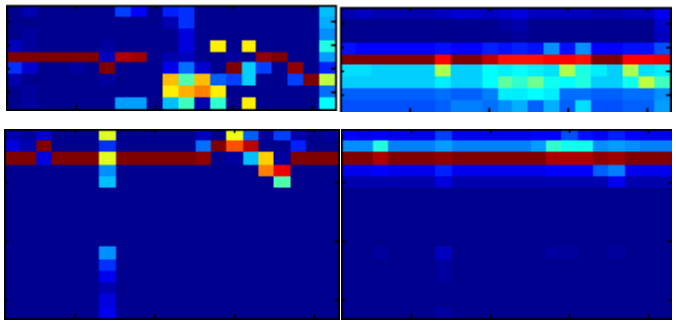
Fig.2 (b), and the correlation-enhance superpixel feature from the identical region in Fig.2 (c), respectively. These superpixels are from the identical semantic region. Ideally, values of their correlation approach to 1. It is seen from Fig.2 (c) that correlation-enhanced superpixels have high correlation. Measure of correlation indicates that the minimum is about 0.1 across original superpixels, and that the minimum is about 0.6 across enhanced superpixels. Visualization and correlation measure show the enhanced superpixels have higher correlation than original superpixels.

Parameters τ and a have an important role in enhancing features. It is very helpful for correlation-enhanced superpixels to enhance the similarity of within-class superpiexls, whereas to reduce the similarity of between-class superpixels.

Finally, LRD optimization Eq.(3) is performed on the correlation-enhanced feature observation X_{new} to get the affinity matrix Z.



(a) Superpixel with label



(b) The original superpixel feature , (c) The correlation-enhanced superpixel feature

Fig.2 original superpixels and correlation-enhanced superpixels are visualized. (a) the image is partitioned into superpixels and marked with label, (b) and (c) illustrate superpixels with index =[17,27,44,52,53,67,69,76,93,97,11,28,24,42,37,40,49,57,55,62,45]. These superpixels are from the identical semantic region. Original superpixels in (b) have poor low-rankness, whereas the enhanced Superpixels in (c) show high correlation, and low-rankness. The enhanced superpixels are very helpful to implement video segmentation.

Our proposed LRD of the correlation-enhanced superpixel for video segmentation is summarized as Algorithm 1.

Algorithm 1 the correlation-enhanced superpixel for video segmentation

Input: each video clip V_i, clip length l, low-rank r.

step1.Partition image frames into superpixels by using SILC;

step2. Extract features and form the superpixel feature observation X ;

step3.Compute pre-defined affinity W and the correlation-enhanced superpixel observation X_{new} ;

step4. Do LRD optimization procedure, **Algorithm 2**,on the X_{new} to get the affinity Z ;

step5. Construct a graph by using (Z+Z^T) as the affinity matrix, apply NCut to this graph to obtain segmentation;

Output: segments result of V_i

C. Optimization

To optimize Eq.(3), we present an Augmented Lagrangian Alternating Direction Method (ALADM) Xu et al. (2017) and formulate

$$L(Z,U,V,E,Y)=\|Z-UV\|_F^2+\lambda\|E\|_1+\frac{\beta}{2}\|X-XZ-E\|_F^2 +\langle Y,X-XZ-E\rangle \quad (6)$$

where β>0 is a penalty parameter, and Y is the Lagrange multiplier corresponding to the constrain $X-XZ-E=0$. We minimize Eq.(6) w.t. U,V,Z,E and iteratively update one variable at a time while fixing the other variables at their latest values, and update the Lagrange multiplier Y and penalty parameter β.

First solving U,V in Eq.(6) is rewritten by

$$L(U,V)=\|Z-UV\|_F^2$$

U,V can be computed by QR factorization. $U = Q$, where Q is the QR factorization of ZV^T. Then we have $V=U^+ Z = Q^T Z$. They are updated by

$$U_{j+1}=QR(Z_jV_j^T),V_{j+1}=U_{j+1}{}^+Z_j \quad (7)$$

where T is transpose, M^+ indicates the pseudo-inverse of the matrix M.

For the solution of Z, via adding $\frac{\beta}{2}\|\beta^{-}Y\|_F^2$, the problem (6) is rewritten as follows:

$$L(Z)=\|Z-UV\|_F^2+\frac{\beta}{2}\left\|X-XZ-E+\frac{Y}{\beta}\right\|_F^2$$

We minimize it and have

$$Z_{j+1}=(\beta X^TX+2I_n)^{-1}(\beta X^T(X-E_j+\frac{Y_j}{\beta})+2U_{j+1}V_{j+1}) \quad (8)$$

For the solution of E in Eq.(6) is rewritten as follows:

$$L(E)=\lambda\|E\|_1+\frac{\beta}{2}\left\|X-XZ-E+\frac{Y}{\beta}\right\|_F^2$$

It is a shrinkage problem, and its closed-form solution is given by

$$E_{j+1}=S_{\lambda/\beta}(X-XZ_{j+1}+\frac{Y_j}{\beta},\frac{\lambda}{\beta}) \quad (9)$$

where $S_\eta(x,\eta)=\mathrm{sgn}(x)\max(|x|-\eta,0)$ is a soft shrinkage operator.

Lagrange multiplier

$$Y_{j+1}=Y_j+\beta(X-XZ_j-E_j) \quad (10)$$

$$\beta=\min(\rho\beta,\overline{\beta}) \quad (11)$$

To summarize the above description given in **Algorithm 2** as following.

Algorithm 2: Optimize Eq.(3) by ALM algorithm

Input: Observation matrix X_{new}, r>0 , $\varepsilon=10^{-8}$, β>0, ρ >0.

Initialization: set V_0,Z_0,Y_0 as zero matrices ;

while *not converged* **do**

update U_{j+1},V_{j+1} by Eq. (7);

update Z_{j+1} by Eq.(8) ;

update E_{j+1} by Eq. (9) ;

update Y_{i+1}, β by Eq. (10~11);

Check the convergence condition

$\|X-XZ_{j+1}-E_{j+1}\|_\infty/\|X\|_\infty<\varepsilon$

end

Output: *Z, U, V, and E*

D. Computational complexity

Now we discuss the time complexity of the proposed Algorithm 1, and Algorithm 2. We propose a correlation-enhanced super-pixel, and perform LRD optimization Algorithm 2 on the enhanced super-pixel observation X_{new}. Comparison with the classical LRD, the added computation is the estimation of predefined affinity W, and the time complexity is $O(n)$. Although Algorithm2 are carried out based on the ALADM framework, its convergence is speeded up significantly owning to the observation X_{new} derived from the linear combination of correlated superpixels. In turn, the running time of the proposed algorithm is reduced.

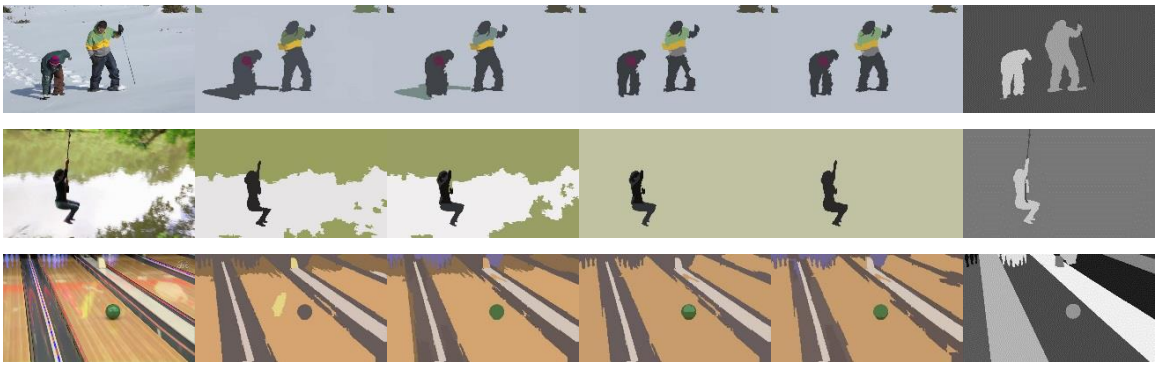
4. Experiments

In this section, our proposed algorithm is evaluated on the standard benchmark VSB100 Galasso et al. (2013) and compared with the state-of-the-art video segmentation algorithms. VSB100 (Video Segmentation Benchmark consisting of 100 HD quality videos) is very challenging, and used for four difficult sub-tasks: general, motion segmentation, non-rigid motion segmentation and camera motion segmentation. Keeping the same setting as Galasso et al. (2013), we regard the general sub-task (60 video sequences) as our test set for all the approaches. The state-of-the-art comparison algorithms: 1) VSS (Video segmentation with superpixels) Galasso et al. (2012), which utilizes twice NCut, first NCut for generating superpixels, second Ncut for grouping superpixels into segments. 2) SPXu, proposed by Xu et al. (2012), is listed at the top in the superpixel-based method, and presents the benchmark and streaming framework. 3) SOLD, presented by Li et al. (2016), which is the classical method in the manner of low-rank optimization.

Parameter setting 1) take the trade-off of time and precision into consideration, the number of superpixel is set to 200. Clip length l is 2~4 frames. Feature dimension lies on the multi-feature cues RGB, HSV, HOG, HOF , $d = n_{rgb} + n_{hsv} + n_{hog} + n_{hof}$. 2) correlation-enhancement $\tau = 0.7$, a is set to 0.1~0.3. 3) optimization processing, fixed rank r=30 and the balance term λ=0.7.

A. *Visualization*

Some of video segmentation samples are visualized. The qualitative comparisons to other algorithms are demonstrated in Fig.3. For the first row of scenarios, the stick region is not treated as one patch when generating superpixels, so all algorithms don't segment it well. For the second row, we set to get 3 segments, the background area is divided into two regions in the algorithms of VSS and SPXu. For the third with severe illumination disturbance, VSS detects the false patch, and SOLD doesn't separate bowling ball from alley border. The shadow is not removed in VSS and SPXu. Compared with columns (d) and (e), our method indicates many detailed regions. One can see that our method qualitatively illustrates the superior performance over the others.



(a)original frame (b) Vss (c)SPXu (d)SOLD (e)our proposed (f) GT

Fig.3 Qualitative comparisons with the state-of-the-art video segmentation methods VSS, SPXu, SOLD.

B. *Empirical Evaluation*

We examine metrics for evaluating both boundary and region benchmark against human ground-truth, and report three different quantities Galasso et al. (2013), for an algorithm: Optimal Dataset Scale (ODS), aggregated at a fixed scale over the dataset, Optimal Segmentation Scale

(OSS), optimally selected for each segmentation, and Average Precision (AP) shown in Tab. 1. It is seen that our proposed algorithm is the superior performance over the others, and AP of BPR is slightly lower than some of the state-of-the-arts. This is due to the case of under-segmentation caused by the inappropriate parameter *a*. We can modify it by developing the approach to generate an appropriate parameter *a.* And we illustrate Boundary precision-recall (BPR) and volume precision-recall (VPR) comparison curves of our proposed algorithm with the state-of-the-art video segmentation approaches: baseline Galasso et al. (2013), VSS, SPXu, SOLD in Fig.4. VSS generates SPs via Ncut, then achieve image segmentation based on SPs. In SPXu the use of steaming framework reduces the consumpiton of memory and space. In SOLD the utilization of low-rank optimization alleviates the influence of the data noise. VSS and SPXu group superpixels into segments based on graph theory. Low-rank decomposition is more effective to mine the affinity than graph theory for the optimization of Low rankness is robust to noise. Comparison with SOLD, our proposed method achieve the better segmentation and high optimization speed owning to enhancing the correlation among the intra-category super-pixels before LRD optimization. Herein, the baseline is much better than others at the cost of more complex image features. From Fig.4 and Tab.1, we can conclude that our approach get the superior performance over the state-of-the-art methods in both BPR and VPR on Dataset VSB10.

Tab.1 Aggregate performance evaluation of boundary precision-recall (BPR) and volume precision-recall (VPR) of state-of-the-art video segmentation algorithms. Bold fonts indicate the best performance.

	BPR			VPR		
Algorithm	ODS	OSS	AP	ODS	OSS	AP
VSS[7]	0.51	0.56	**0.45**	0.45	0.51	0.42
SPXu[6]	0.38	0.46	0.32	0.45	0.48	0.44
SOLD[17]	0.54	0.58	0.40	0.53	0.60	0.46
ESP (Our)	**0.54**	**0.60**	0.42	**0.55**	**0.62**	**0.48**

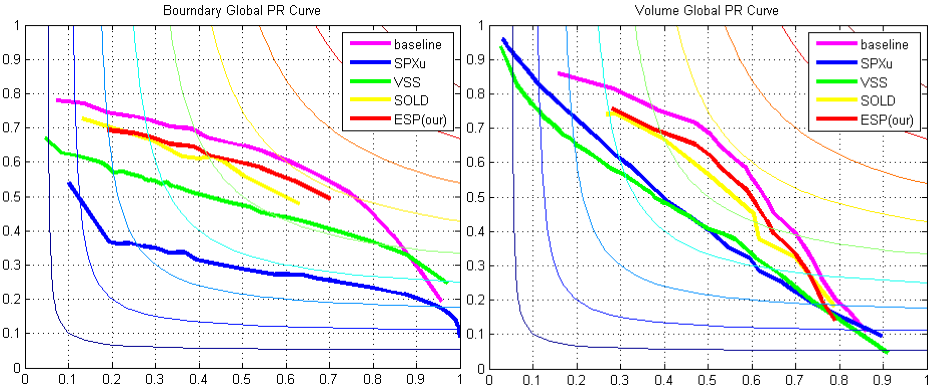


Fig. 4 Boundary precision-recall (BPR) and volume precision-recall (VPR) comparison curves of our proposed algorithm with the state-of-the-art video segmentation approaches VSS, SPXu, SOLD.

C. *Efficiency analysis*

To test the effectiveness of the correlation-enhanced super-pixel, we vary the controlling parameter of the correlative component *a* at a range of 0~0.5. All experiments are conducted on a notebook computer with Core i5CPU at 2.2GHz and 6GB of memory, running windows7 and Matlab R2014a. Running time comparisons are listed in Tab.2 when performing the LRD optimization on the original superpixel (LRD-SP) and the correlation-enhanced superpixel (LRD-ESP), respectively. The convergence curves are illustrated in Fig.5. It is seen from Fig.5 and Tab.2 that the correlation-enhanced superpixel is optimized with high efficiency.

Tab. 2 Comparison of run timing

	LRD-SP	LRD-ESP
Average Iteration Number	16	5
Running time per frame (second)	1.67s	0.2s

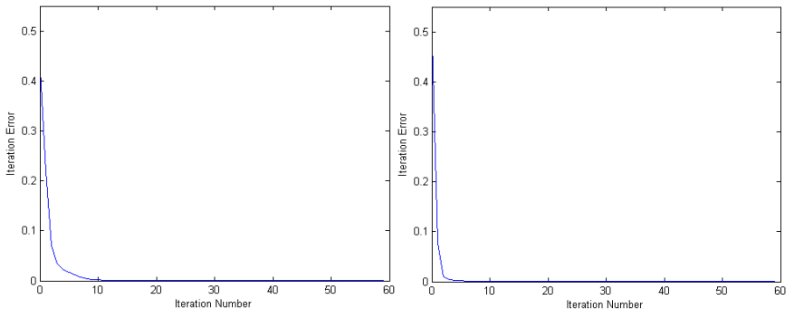


Fig.5 (a) LRD- SP (b) LRD- ESP

5. Conclusion

This paper proposes a correlation-enhanced superpixel for video segmentation in the framework of low-rank decomposition, and evaluates the performance on Dataset VSB100. Our contributions are 1) constructing the correlation-enhanced superpixel. We enhance the correlation of superpixels via the linear combination of superpixels. It is very helpful to enhance the affinity of within-class superpixels, whereas to reduce the correlation of between-class superpixels. 2) designing the optimization algorithm of LRD. The optimization of LRD performed on the enhanced superpixels speeds up. 3) achieving video segmentation owning to the affinity and high correlation superpixels. It is very feasible for the correlation-enhanced super-pixel to perform LRD for video segmentation. The future work is to modify the approach to establish the initial affinity W.

Acknowledgements This work was supported by National Natural Science Foundation of China (No 61602397) , The Natural Science Foundation of Hunan Province(2017JJ2251,2017JJ3315), and Chinese Scholarship Council of the Ministry of Education .

Compliance with ethical standards

Conflict of interest Author Haixia Xu declares that he has no conflict of interest. Author Edwin R. Hancock declares that he has no conflict of interest. Author Wei Zhou declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Felzenszwalb P, Huttenlocher D (2004) Efficient Graph-Based Image Segmentation. International Journal of Computer Vision, 59(2):167-18

Grundmann M, Kwatra V, Han M, et al (2010) Efficient hierarchical graph-based video segmentation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2141-2148

Shi J, Malik J (2000) Normalized Cuts and Image Segmentation, IEEE Transactions on PAMI, 22 (8): 888-905

Wang Y, Jiang Y, Wu Y, et al (2011) Spectral clustering on multiple manifolds. IEEE Transactions on Neural Networks, 22(7):1149-1161

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2010) Slic superpixels. In Technical report, *EPFL*

Galasso F, Cipolla R, Schiele B (2012) Video segmentation with superpixels. In Proc. Asian Conf. Computer Vision, 760-774

Xu C, Corso J (2012). Evaluation of super-voxel methods for early video processing. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, doi: 10.1109/CVPR.2012.6247802

Galasso N, Nagaraja J, Cardenas T, Brox B, Schiele A (2013) Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis, International Conference on Computer Vision, doi:10.1109/ICCV.2013.438

Liu G, Lin Z, Yan S, et al (2013). Robust Recovery of Subspace Structures by Low-Rank Representation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 35(1):171-184

Yin M, Gao J, Lin Z (2016). Laplacian Regularized Low-Rank Representation and Its Applications. IEEE Transactions on Pattern Analysis & Machine Intelligence, 38(3):504-517

Zhang T, Ghanem B, Liu S, et al (2013) Low-Rank Sparse

Coding for Image Classification. In Proceedings of IEEE International Conference on Computer Vision, 281-288

Cheng B, Liu G, Wang J, et al (2011). Multi-task low-rank affinity pursuit for image segmentation. In Proceedings of IEEE International Conference on Computer Vision, 2439-2446

Wang L, Dong M (2012) Multi-level Low-rank Approximation based Spectral Clustering for image segmentation. Pattern Recognition Letters, 33(16):2206-2215

Li T, Bin Cheng B,Ni B et al (2016a) Multitask Low-Rank Affinity Graph for Image Segmentation and Image Annotation. Acm Transactions on Intelligent Systems & Technology, 7(4):1-18

Li C, Lin L, Zuo W, Wang W, Tang J (2016b) An Approach to Streaming Video Segmentation with Sub-Optimal Low-Rank Decomposition. IEEE Transactions on Image Procssing (T-IP), 25(5):1947-1960

Li C, Lin L, Zuo W, Wang W, Tang J ,Yan S (2015) SOLD: Sub-Optimal Low-Rank Decomposition for Efficient Video Segmentation, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 5519-5527. doi:10.1109/CVPR.2015.7299191

Shelhamer E, Long J, Darrell T (2017) Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):640-651

Farnoush Z，Borislav A，Jan S (2018). Superpixel-based Road Segmentation for Real-time Systems using CNN, In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), 257-265

Das A , Ghosh S , Sarkhel R , et al.(2018) Combining Multi-level Contexts of Superpixel using Convolutional Neural Networks to perform Natural Scene Labeling, arXiv:1803.05200v1

Liu R, Lin Z, Torre F, Su Z (2012) Fixed-Rank Representation for Unsupervised Visual Learning, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 598-605

Brox T, Malik J (2010) Object segmentation by long term analysis of point trajectories. In Proceedings of European Conference on Computer Vision, doi：10.1007/978-3-642-15555-0_21

Wen Z, Yin W, Zhang Y (2010) Solving a Low-Rank Factorization Model for Matrix Completion by a Non-linear Successive Over-Relaxation Algorithm, Rice CAAM Tech Report TR10-07.

Xu H, Zhou W, Wang Y, Wang W, Mo Y(2017) Matrix Separation Based on LMaFit-Seed. The Computer Journal, 60(11):1609-1618

Duta I, Uijlings J, Nguyen T, et al (2016) Histograms of Motion Gradients for real-time video classification. International Workshop on Content-Based Multimedia Indexing, doi: 10.1109/CBMI.2016.7500260

Konstantinos G (2008) The Bhattacharyya. Measure, Version 1.0, March 20

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 886-893

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, arXiv:1512.03385, doi:10.1109/CVPR.2016.90

Chen L, Zhu Y, Papandreou G, et al (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, preprint arxiv: 1802.02611

Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmenta-tion. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 9351: 234-241

Chen L, Papandreou G, Kokkinos I, Murphy K et al. (2016). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis & Machine Intelligence, 40(4):834-848.

Luc P, Couprie C, Chintala S, et al (2016) Semantic Segmentation using Adversarial Networks, NIPS-2016 NIPS Workshop on Adversarial Training, Barcelona, Spain, arXiv:1611.08408